Academy @

Stamicarbon

Intro to Data Science

▸ Day 1 → 5th Sep (13:00-17:00):
An Overview of Data Science Technologies

▸ Day 2 → 12th Sep (10:00-15:00):
Modern Applications of Data Science Methods

Trainer: Pedro V Hernández Serrano

Academy @ Stamicarbon

Intro to Data Science

▸ Day 1 → 5th Sep (13:00-17:00):
An Overview of Data Science Technologies

▸ Day 2 → 12th Sep (10:00-15:00):
Modern Applications of Data Science Methods

Trainer: Pedro V Hernández Serrano
< Data Scientist - Maastricht University >

Format

- 1 Hour lecture (incl. Q&A)
- 20 mins. Set up
- 40 mins. Live examples (Part 1)
- 20 mins. Questionnaire + Break
- 40 mins. Live examples (Part 2)
- 1 Hour hands-on

An Overview of Data Science Technologies

Pedro V Hernández Serrano
Workshop (Day 1) - Academy @Stamicarbon
05/09/2022

Stamicarbon

Trainer

Data Scientist
- Certified Cloud Solutions Engineer (GCP)
- Certified Data Protection Officer (ECPC)
- MSc Data Science, BSc Actuarial Science

Current
- Faculty **Data Steward** at Maastricht University
- **Cloud Architect** at European Trade Union Institute (ETUI)
- **Research** Fellow at Netherlands eScience Center
- **Freelance** Data Science & Data Governance Bootcamps

Contact:
Site: pedrohserrano.com
LinkedIn: linkedin.com/in/pedrohserrano/
Twitter: @pedrohserrano

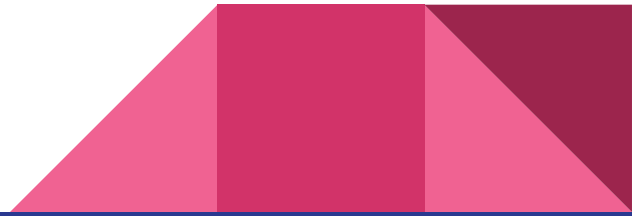
Lecture

Day 1

- What is Data Science
- Data Science: Technologies vs Methods
- Applications
  - EDA
  - Cloud Services
  - Data Annotation
  - Knowledge Databases
- Data Science with Python

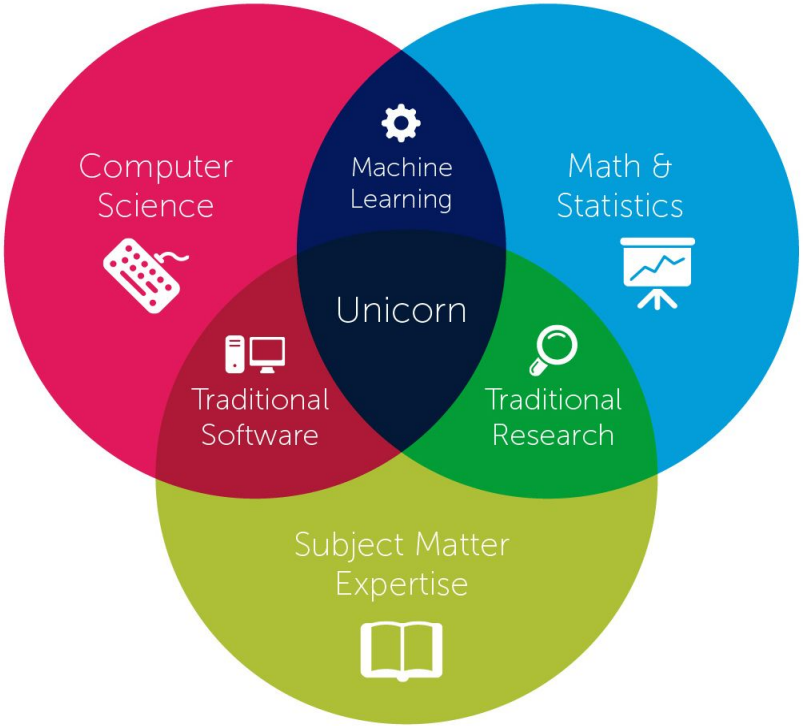
Goals

- Historical recap and context around data science
- Explore some main technologies used in data science
- Introduce the basics of the python programming environment.
- Develop a general understanding of data formats and representations
- Get an overview of some python visualization packages.
- Know limitations and caveats of available interactive python libraries
- We will use open source analytics tools
- We will learn through doing.



What is Data Science?

What is Data Science



Homes

Smart homes will help you save on utility bills and help the electric company reduce the grid's impact on the environment.

Crime Prevention

Data is helping police forces understand where and when crimes are most likely to happen. Helping them keep us safe.

Government

Governments use city data to adapt and improve public services such as transport, parking, garbage disposal, etc. Making your community a pleasure to live in.

Emergency Services

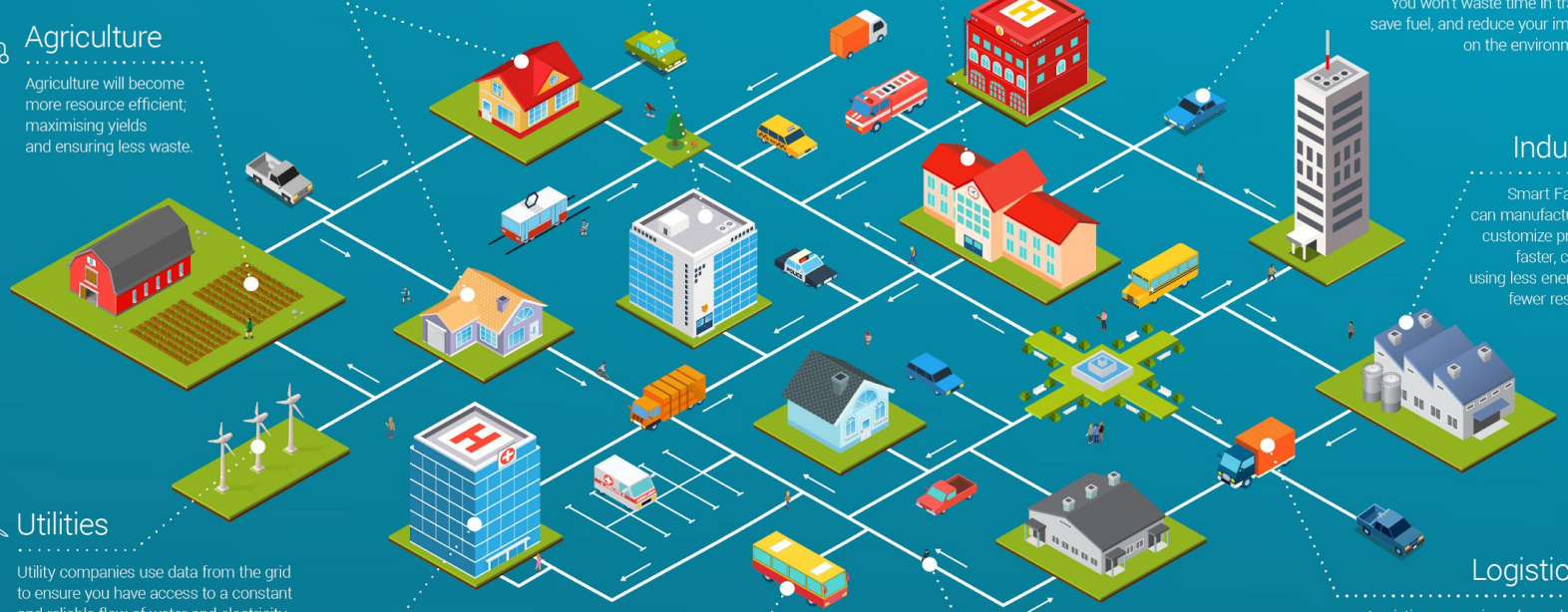
Data is helping emergency services mobilise response teams to where they are needed most; increasing the odds of successful rescue missions.

Traffic

Cars will share GPS & speed data making traffic jams a thing of the past. You won't waste time in traffic, save fuel, and reduce your impact on the environment.

Agriculture

Agriculture will become more resource efficient; maximising yields and ensuring less waste.



Industry

Smart Factories can manufacture and customize products faster, cheaper, using less energy and fewer resources

Utilities

Utility companies use data from the grid to ensure you have access to a constant and reliable flow of water and electricity.

Logistics

Logistic companies rely on data and algorithms to make the most efficient use of transport systems. This not only reduces their costs, but also their carbon emissions.

Healthcare

Medical researchers will be able to determine what affects diseases; providing better detection and treatment

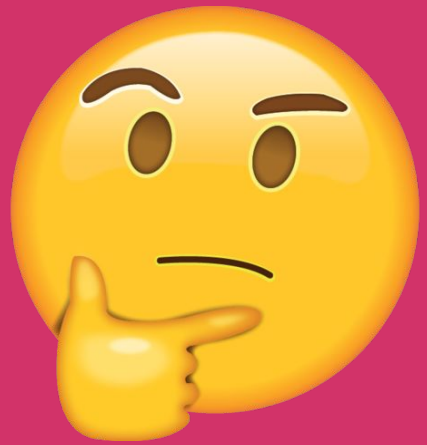
Public Transport

Public transport will take into account traffic conditions and peak usage to give you better recommendations and get you to your destination faster.

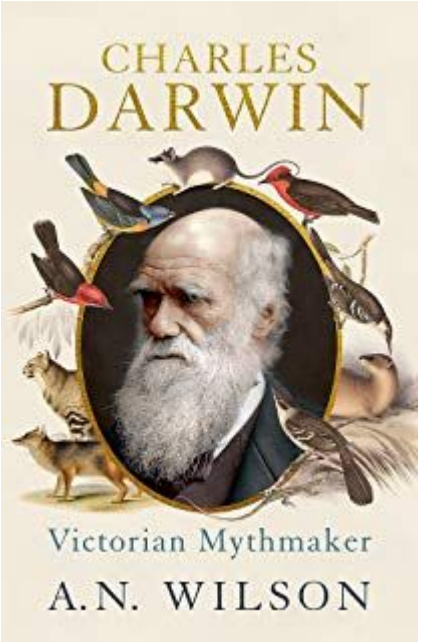
Connectivity

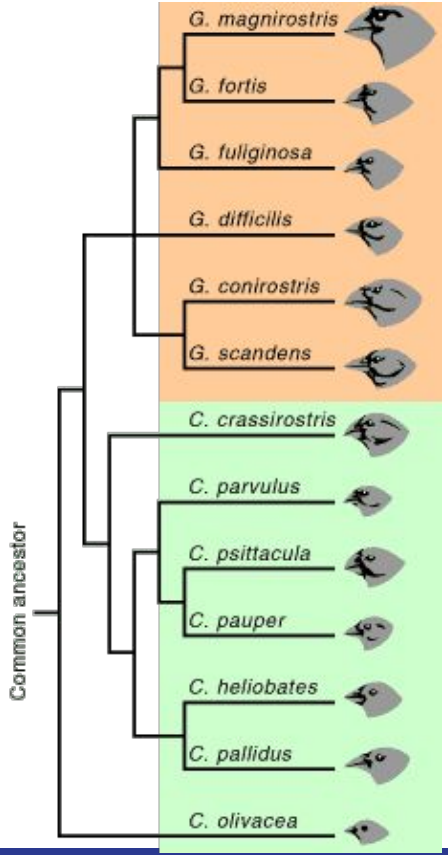
Our devices are enabling us to access and connect to many services backed by data; improving the way we live, work and play.

Is Data Science a new thing?



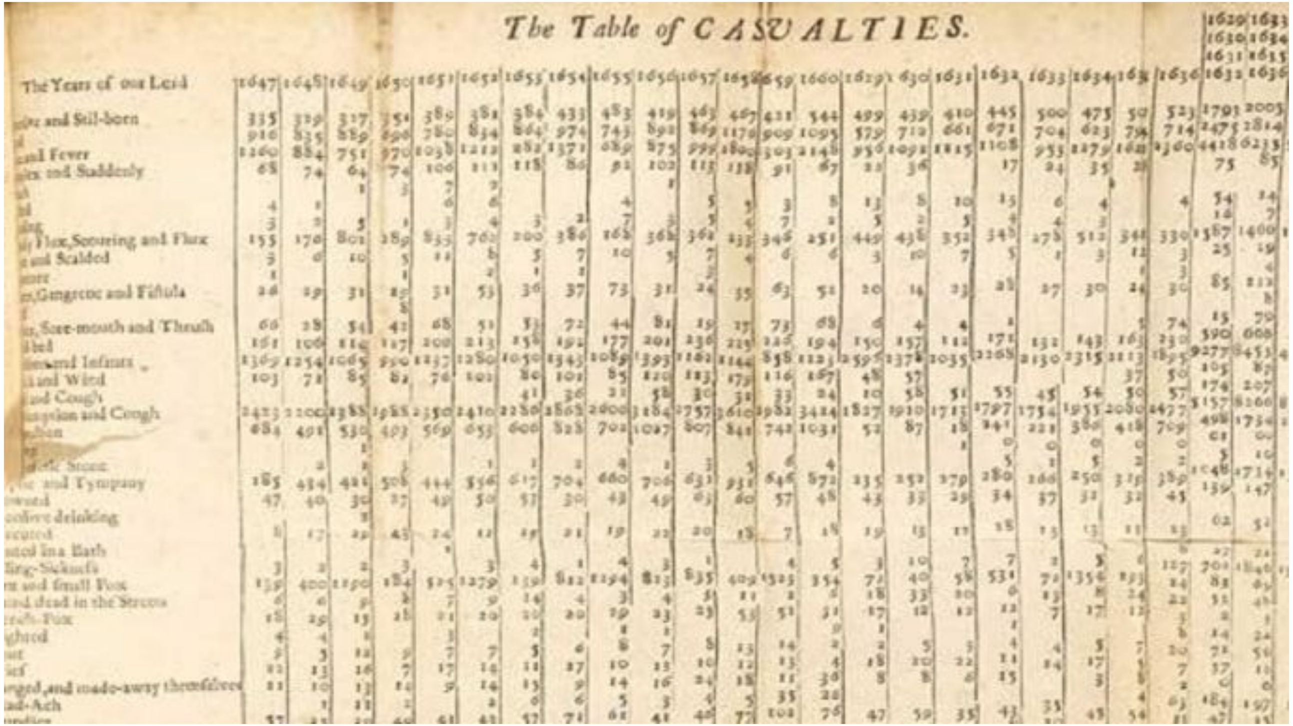
Observation and categorization





The reality is that we stored information for analytical purposes for quite a while
Many people in academia say that every scientist is a data scientist.

Statistics early stage



18th century:

Systematic collection of demographic and economic data by states.

For tax or military use...

Smithsonian Institute Archives: siarchives.si.edu/blog/link-love-4-28-2017

So for decades statisticians, economists, demographers, were very happy doing data analysis, till...

The PC happened

No manual analysis anymore

thedayintech.wordpress.com/2013/03/08/4995-00-for-10-megabytes/

Then new scientists, chemists, econometricians were very happy doing data analysis, till...

The Big Data happened

We needed to study the technologies to harness data

The Data Science profession was born

The Data Scientist aimed to

use “new” **technologies**

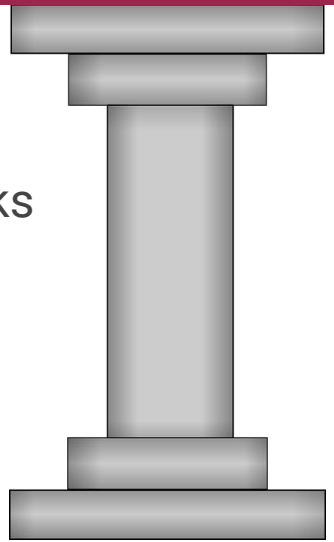
to adapt “old” **methods**

Data Science: Technologies vs Methods

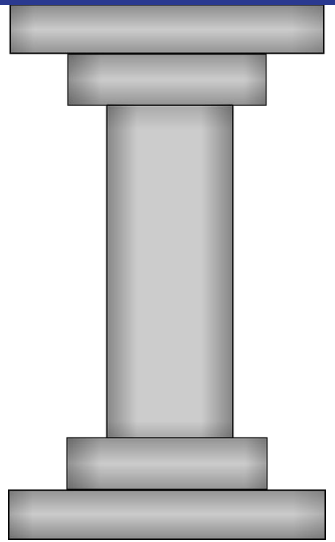
Data Science pillars

Technologies

- Python
- Jupyter Notebooks
- Neo4j
- Tableau
- SQL



Methods

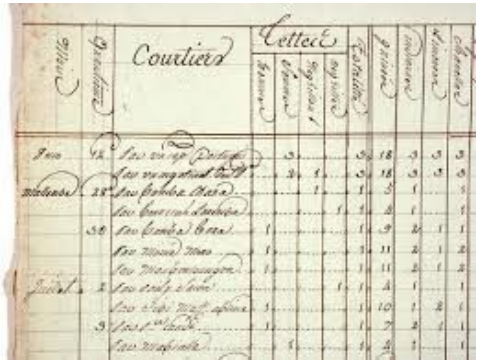


- EDA
- Machine Learning
- NLP
- Network Analysis

Technologies

Technologies

Manual



User oriented

Office

IBM SPSS

Purpose oriented

tableau

neo4j

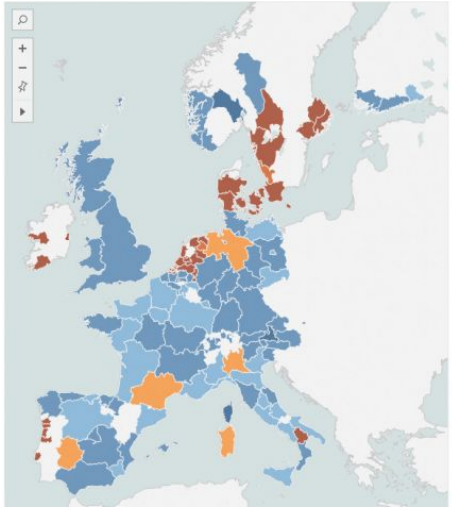
Software

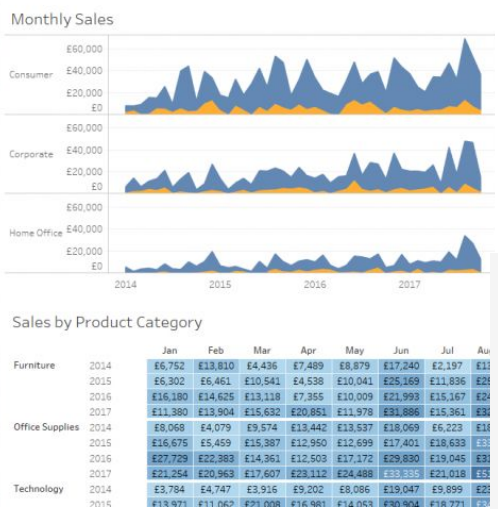
Programing Language

R

python™

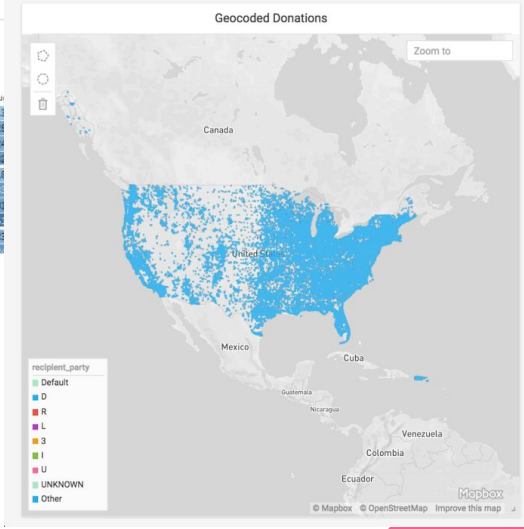
Purpose Oriented Technologies

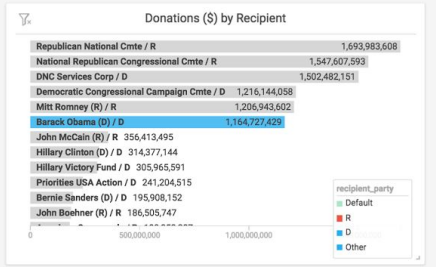


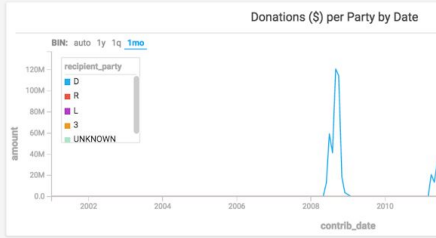


Tableau

Python

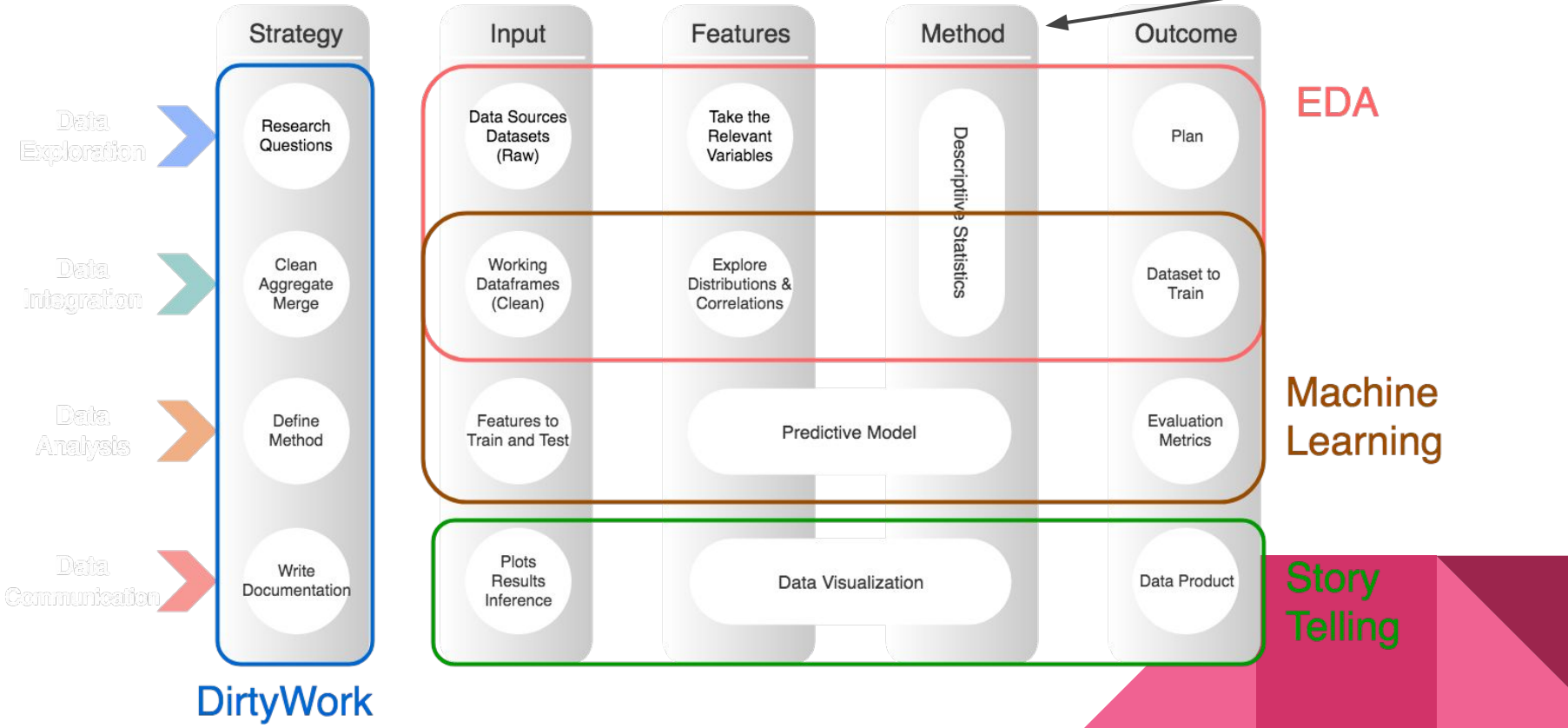






Methods

Methods in Data Science



Nutshell

Technologies in this workshop

- Python + Notebooks
- Neo4j

Methods in this workshop

- EDA
- Machine Learning

Applications

Applications

Technologies + Methods

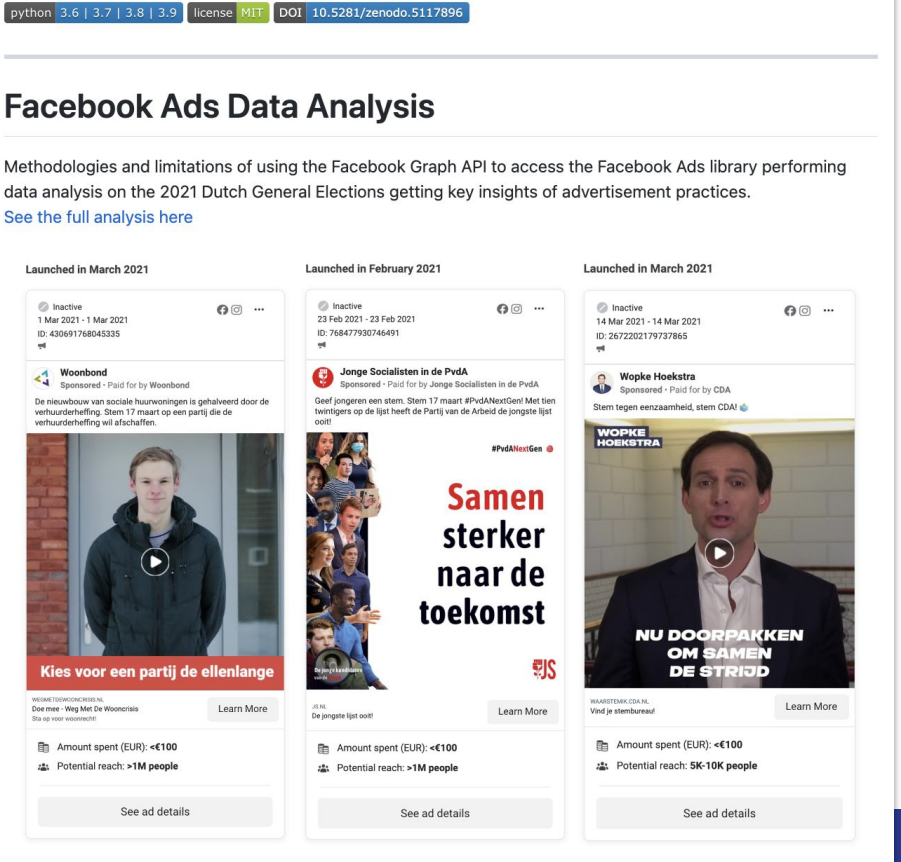
EDA

Simulations

Data Annotation

Knowledge Database

Exploratory Data Analysis (EDA)



Simulation

Create a synthetic dataset of workers

```
import crowded.simulate as cs

#define your parameters
total_workers = 40
alpha = 28
beta = 2
#create task dataset
df_workers = cs.Workers(alpha, beta).create(total_workers)
```

Assign easily and fairly workers to tasks

```
import crowded.simulate as cs

#workers per task should always be smaller than the number of workers
wpt = 5
#create assignment
df_tw = cs.AssignTasks(df_tasks, df_workers, wpt).create()
```

Compute Bayes probability and predict worker answers

```
import crowded.method as cs

#workers per task should always be smaller than the number of workers
wpt = 5
#create assignment
df_tw = cs.AssignTasks(df_tasks, df_workers, wpt).create()
```

Compute Bayes probability and Predict answers of the workers

```
import crowded.method as cm

#define the parameters
x = df_tw['prob_task'] #vector of probabilities of tasks
y = df_tw['prob_worker'] #vector of probabilities of workers
z = df_tasks['true_answers'].unique()  #vector of valid answers in the experiment
#compute probability
cp = cm.ComputeProbability(x, y, z)
```

RESEARCH-ARTICLE

CrowdED: Guideline for Optimal Crowdsourcing Experimental Design

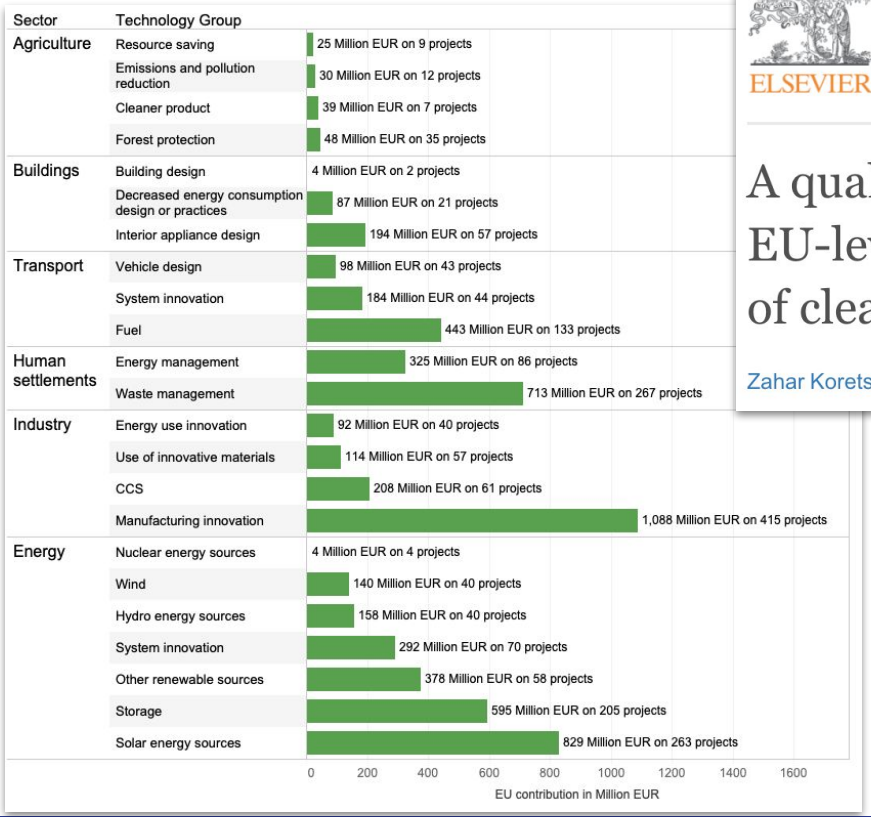
Authors: Amrapali Zaveri, Pedro Hernandez Serrano, Manisha Desai, Michel Dumontier

Authors Info & Affiliations

Publication: WWW '18: Companion Proceedings of the The Web Conference 2018 • April 2018 • Pages 1109–1116 • https://doi.org/10.1145/3184558.3191543

CrowdED: Guideline for Optimal Crowdsourcing Experimental Design
A Zaveri, Pedro Hernandez Serrano, M Desai, M Dumontier
Companion of the The Web Conference 2018 on The Web Conference 2018, 1109-1116

Data Annotation



ELSEVIER

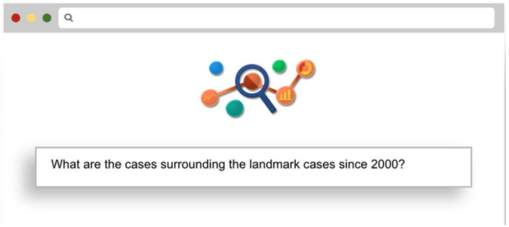
Current Research in Environmental Sustainability

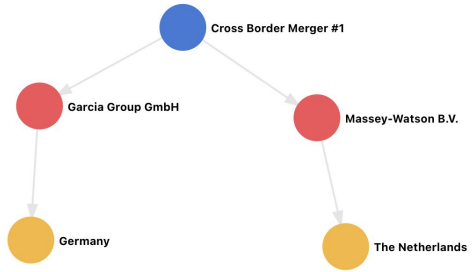
Volume 3, 2021, 100084

A qualitative-computational cataloguing of the EU-level public research and innovation portfolio of clean energy technologies (2014–2020)

Zahar Koretsky [a], Pedro V. Hernández Serrano [b], Seun Adekunle [b], Michel Dumontier [b]

Knowledge Databases





The Case for a Linked Data Research Engine for Legal Scholars

Published online by Cambridge University Press: **04 November 2019**

Kody MOODLEY, Pedro V HERNANDEZ-SERRANO, Amrapali J ZAVERI, Marcel GH SCHAPER, Michel DUMONTIER and Gijs VAN DIJCK

https://eu-corporate-mobility.org/

Exploratory Data Analysis (EDA)

from Data to Viz

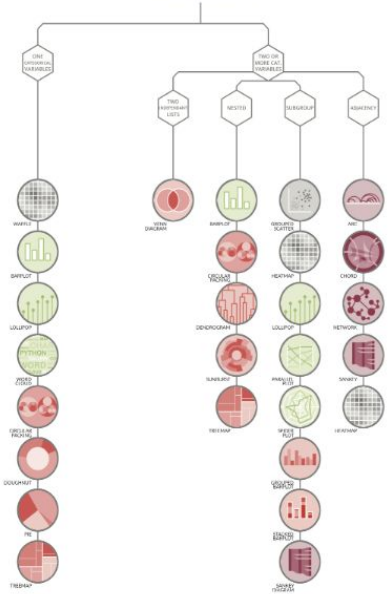
'From Data to Viz' is a classification of chart types based on input data format. It will help you find the perfect chart in three simple steps :

1. Identify what type of data you have.
2. Go to the corresponding decision tree and follow it down to a set of possible charts.
3. Choose the chart from the set that will suit your data and your needs best.

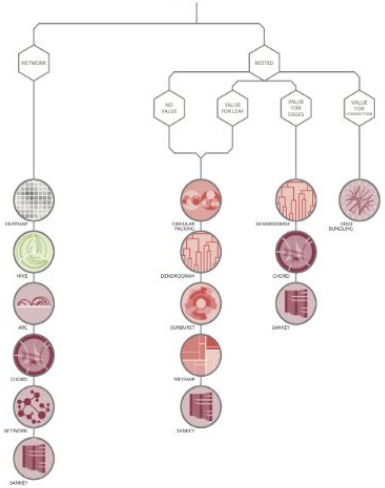
Dataviz is a world with endless possibilities and this project does not claim to be exhaustive. However it should provide you with a good starting point. For an interactive version and much more, visit:

data-to-viz.com

CATEGORIC

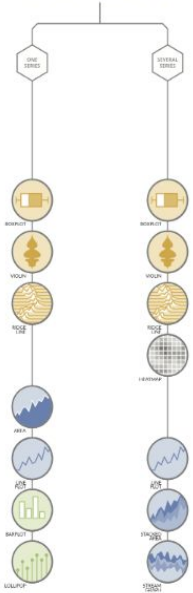


RELATIONAL



MAP

TIME SERIES

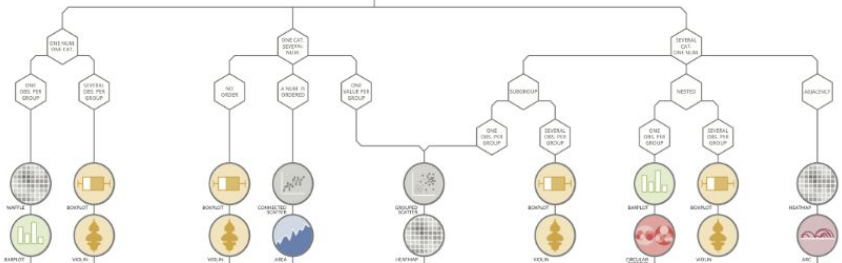


WHAT DO YOU WANT TO SHOW ?

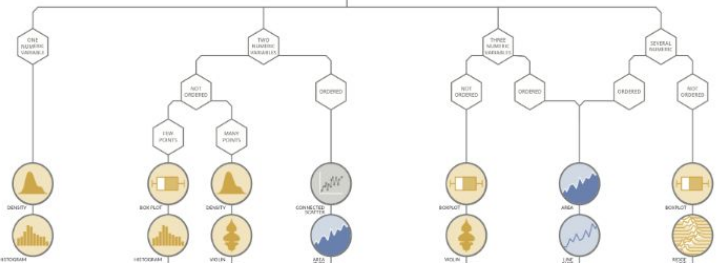
- Distribution
- Correlation
- Ranking
- Part of a whole
- Evolution
- Maps
- Flow

data-to-viz.com/

CATEGORIC AND NUMERIC



NUMERIC



Technologies for EDA

Proprietary Software

- Shorter learning curve
- Automatic updates
- Specific formats
- Enterprise licence

Open Source

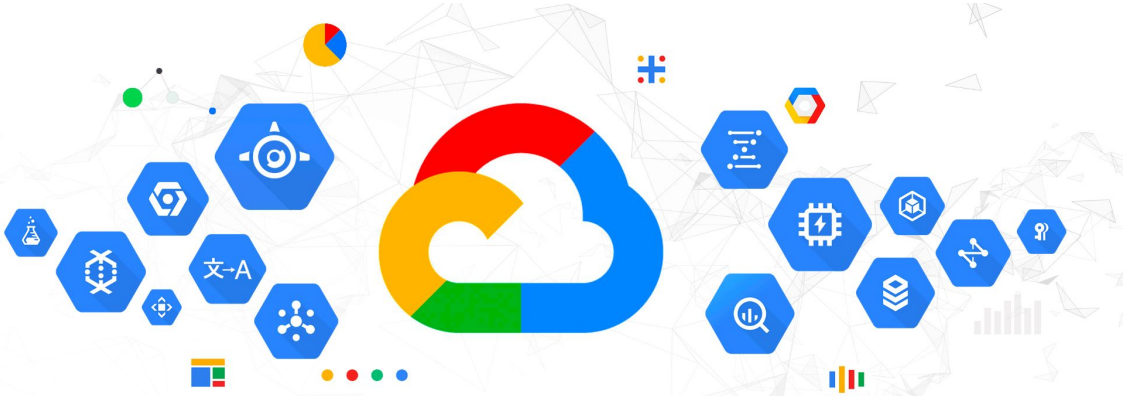
- Longer learning curve
- Harder Maintenance
- Interoperable
- Free

Cloud Services

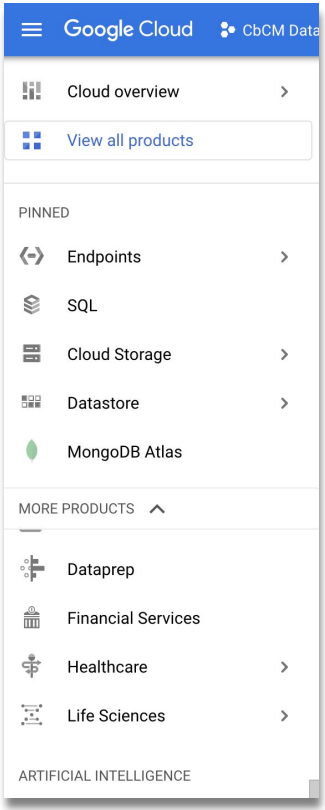
Cloud Services

- IaaS (Infrastructure as a service)
- Pay as you go business model
- Harmonized data centers aiming Net 0% CO2

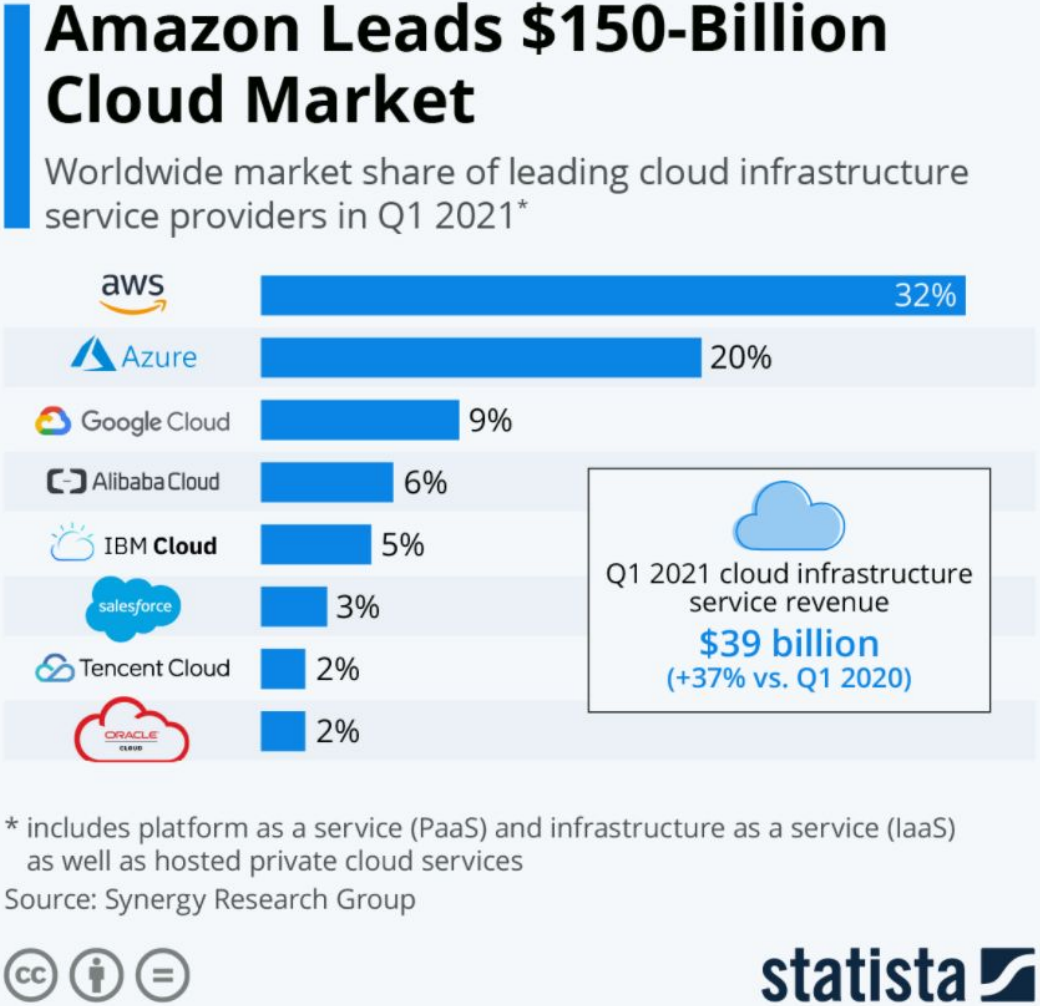
- Big data simulations should move out of your computer
- One-stop shop for all data problems
- Governance models & security



Cloud Services

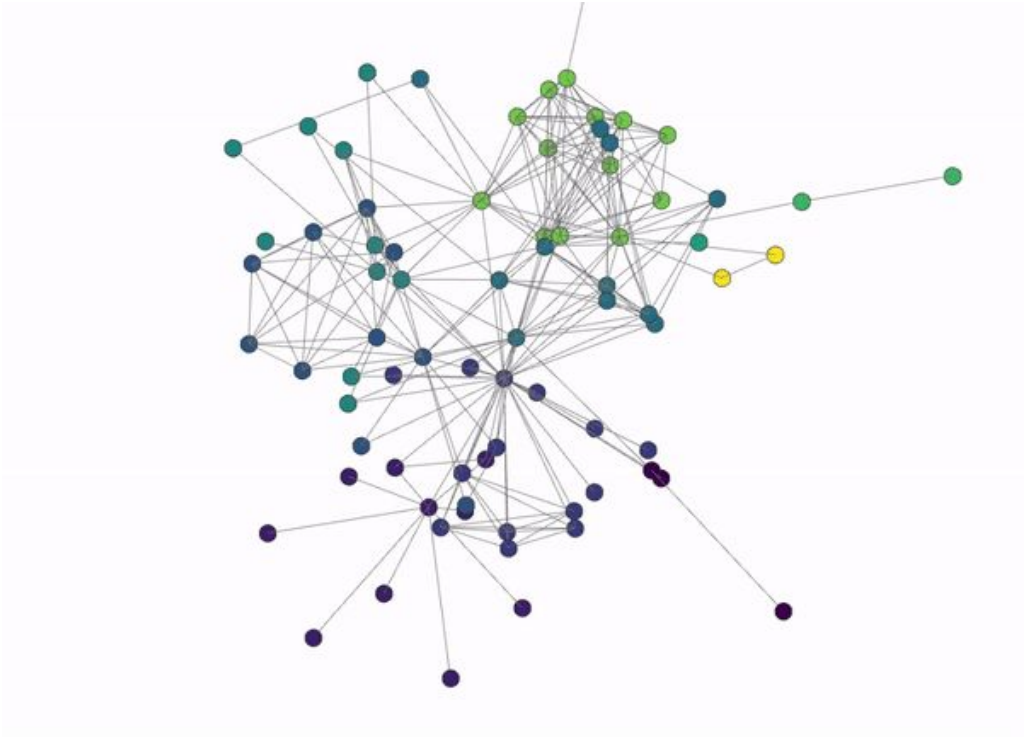


All services in one console



Understanding Data Annotation

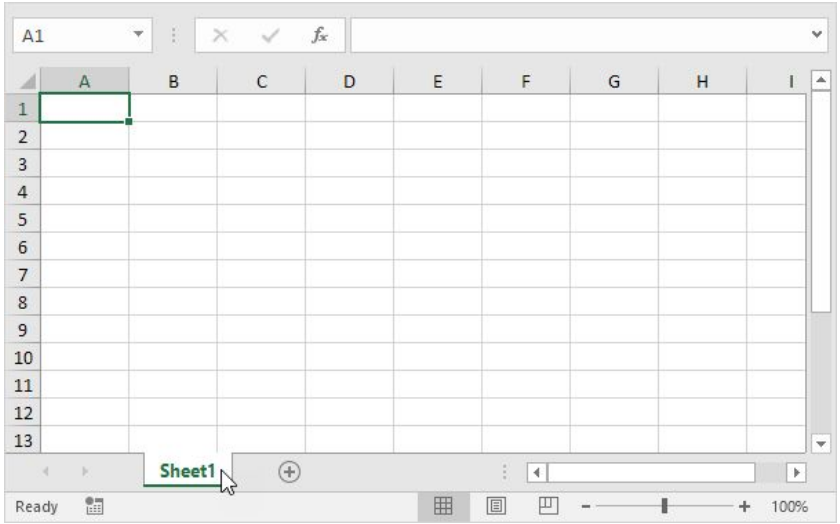
Data Stuff…



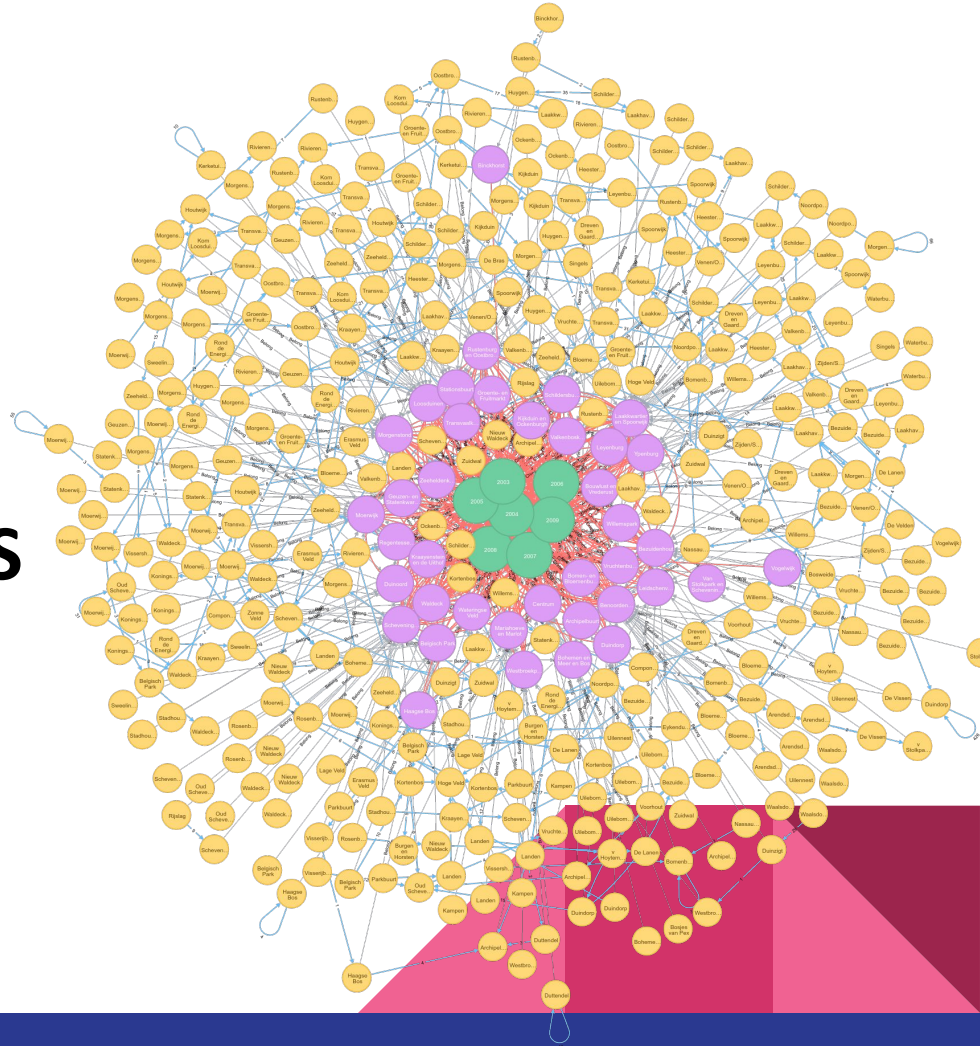
- Data Model
- Database
- Dataset
- Data Syntax
- **Data Format**

But before we can annotate and enhance data, we need to understand the terminology and especially the data formats

Data Format



VS



Tables

Columns

Rows

Country	Name	Year Passed	Executiv...	Fra...	Categories	Document
Japan	Act on the Improvem...	01/01/2015	Legislative	*null*	Energy De...	Law
Macedoni...	Action Plan on Rene...	01/01/2015	Executive	*null*	Energy Sup...	Plan
Czech Rep...	Adaptation strategy ...	01/01/2015	Executive	Adapt...	Adaptation	Strategy
Niue	Agriculture Sector Pl...	01/01/2015	Executive	*null*	Adaptation...	Plan
Thailand	Alternative Energy D...	01/01/2015	Executive	*null*	Energy Sup...	Plan
United Sta...	Clean Power Plan	01/01/2015	Executive	*null*	Energy Sup...	Plan
Malta	Climate Action Act	01/01/2015	Legislative	Mitig...	Adaptation...	Law
Ireland	Climate Action and L...	01/01/2015	Legislative	Mitig...	Adaptation...	Law
Ireland	Climate Action and L...	01/01/2015	Legislative	Mitig...	Adaptation...	Law

Tables

climate_change... Country	Abc climate_change_laws.csv Name	climate_change_la... Year Passed	Abc climate_change_l... Executiv...	Abc climate_c... Fra...	Abc climate_change_l... Categories	Abc climate_change Document...
Japan	Act on the Improvem...	01/01/2015	Legislative	*null*	Energy De...	Law
Macedoni...	Action Plan on Rene...	01/01/2015	Executive	*null*	Energy Sup...	Plan
Czech Rep...	Adaptation strategy ...	01/01/2015	Executive	Adapt...	Adaptation	Strategy
Niue	Agriculture Sector Pl...	01/01/2015	Executive	*null*	Adaptation...	Plan
Thailand	Alternative Energy D...	01/01/2015	Executive	*null*	Energy Sup...	Plan
United Sta...	Clean Power Plan	01/01/2015	Executive	*null*	Energy Sup...	Plan
Malta	Climate Action Act	01/01/2015	Legislative	Mitig...	Adaptation...	Law
Ireland	Climate Action and L...	01/01/2015	Legislative	Mitig...	Adaptation...	Law
Ireland	Climate Action and L...	01/01/2015	Legislative	Mitig...	Adaptation...	Law

.xls

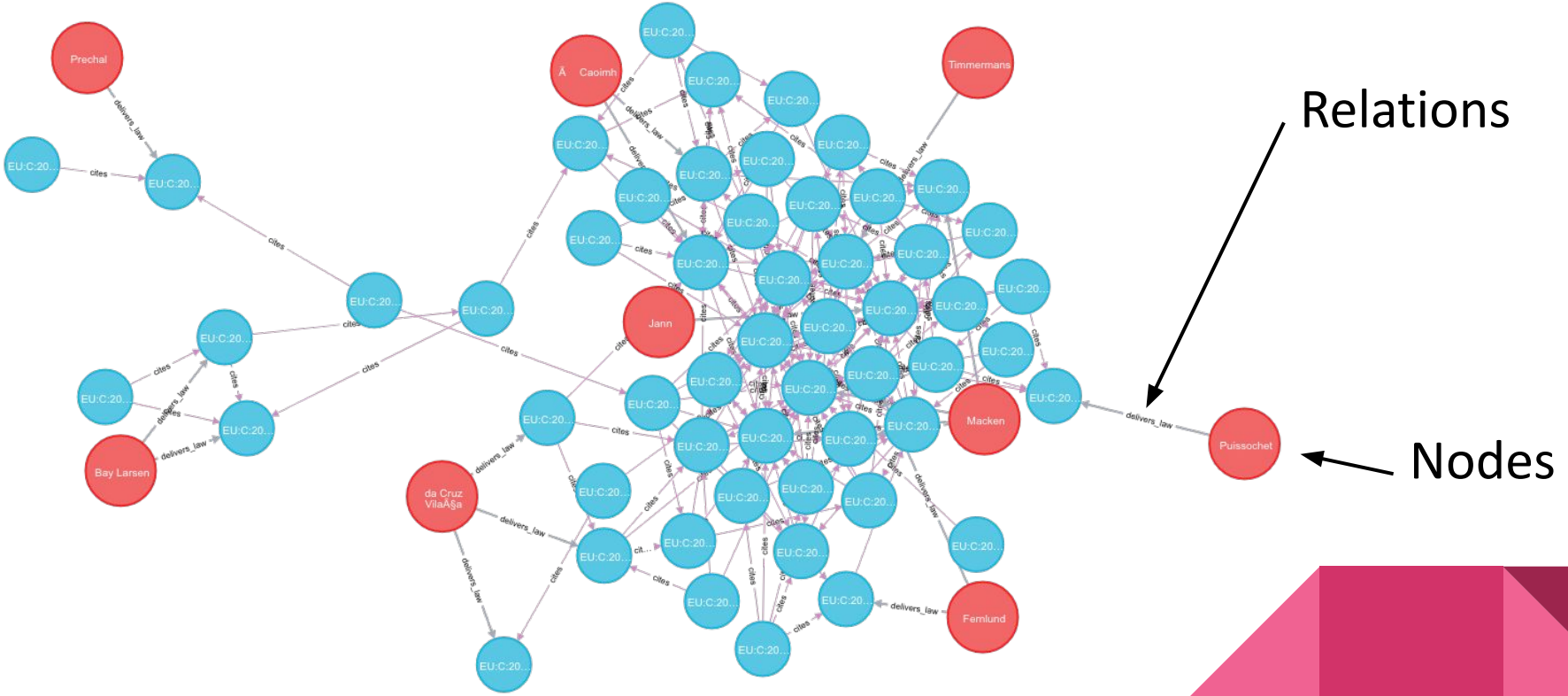
.csv

.tsv

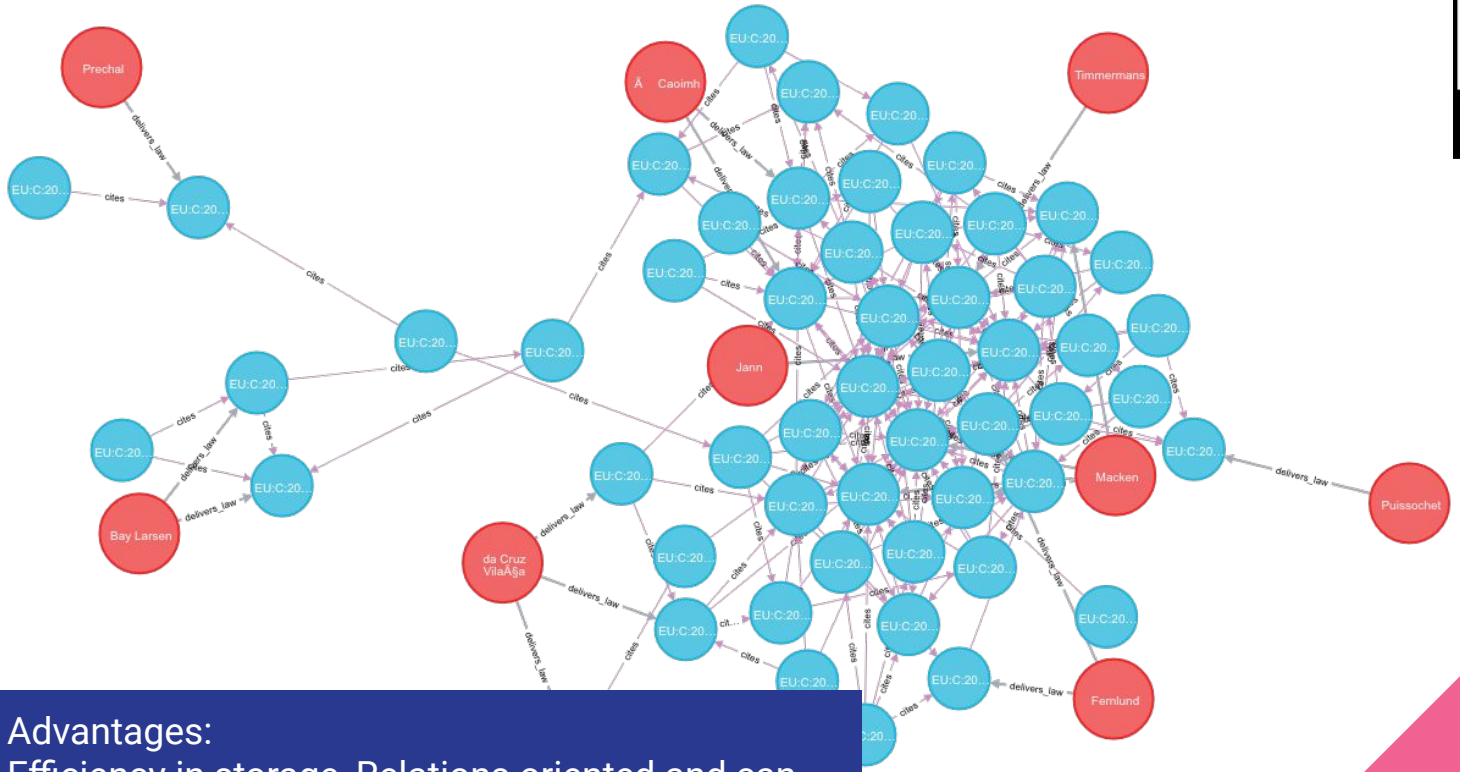
.data

Advantages:
File sizes and compatibility with other systems

Graphs


Prechal
Caoimh
Timmermans
Jann
Macken
Puissochet
Bay Larsen
da Cruz
Femlund
cites
delivers_law
Relations
Nodes

Graphs



.nt

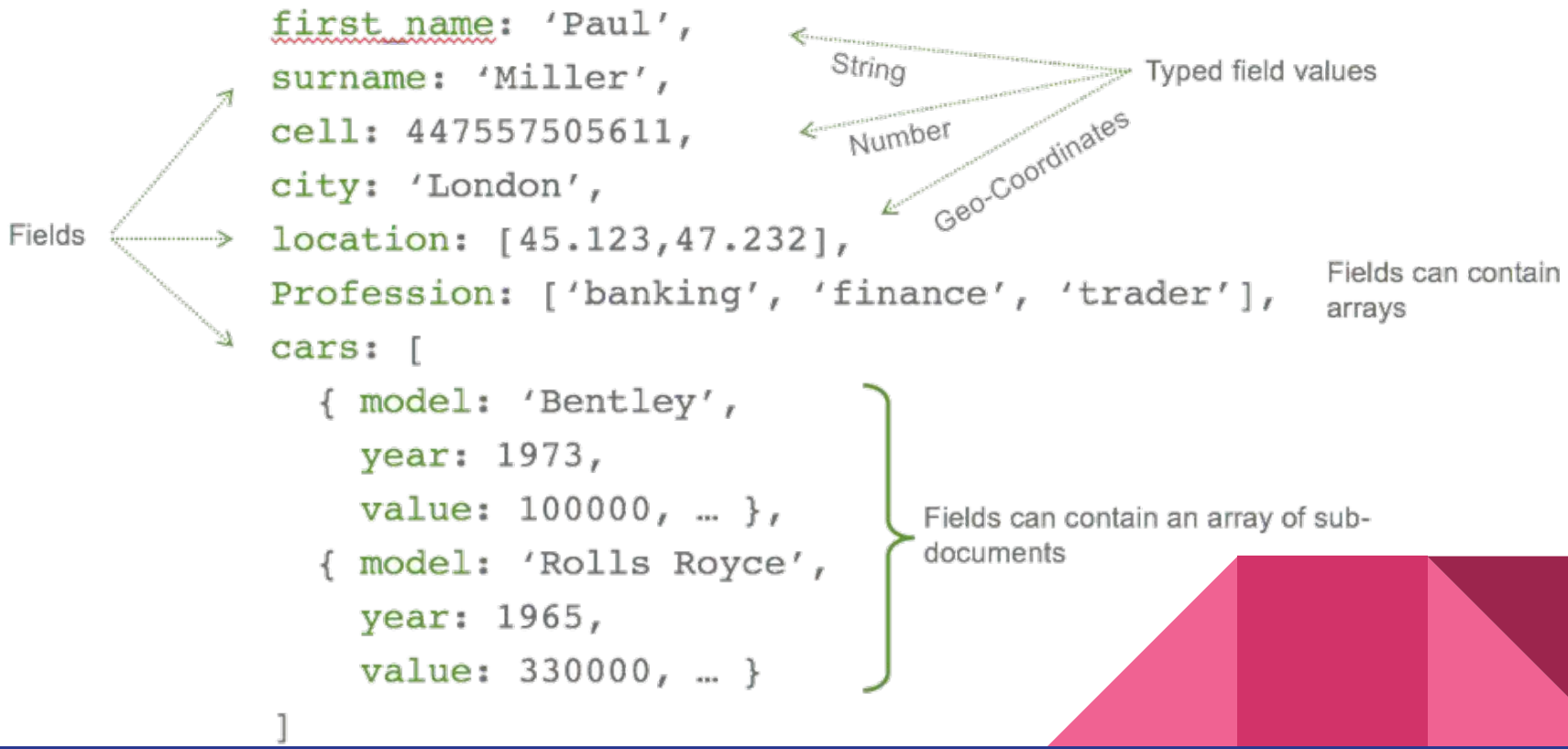
.ttl

.gexf

.grapML

Advantages:
Efficiency in storage, Relations oriented and can be easily integrated with external data sources

Documents


first_name: 'Paul',
surname: 'Miller',
cell: 447557505611,
city: 'London',
location: [45.123,47.232],
Profession: ['banking', 'finance', 'trader'],
cars: [
{ model: 'Bentley',
year: 1973,
value: 100000, … },
{ model: 'Rolls Royce',
year: 1965,
value: 330000, … }
]
Fields
String
Number
Geo-Coordinates
Typed field values
Fields can contain arrays
Fields can contain an array of sub-documents

Documents

```
first_name: 'Paul',
surname: 'Miller',
cell: 447557505611,
city: 'London',
location: [45.123,47.232],
Profession: ['banking', 'finance',
cars: [
  { model: 'Bentley',
    year: 1973,
    value: 100000, … },
  { model: 'Rolls Royce',

                        , … }
```

.json

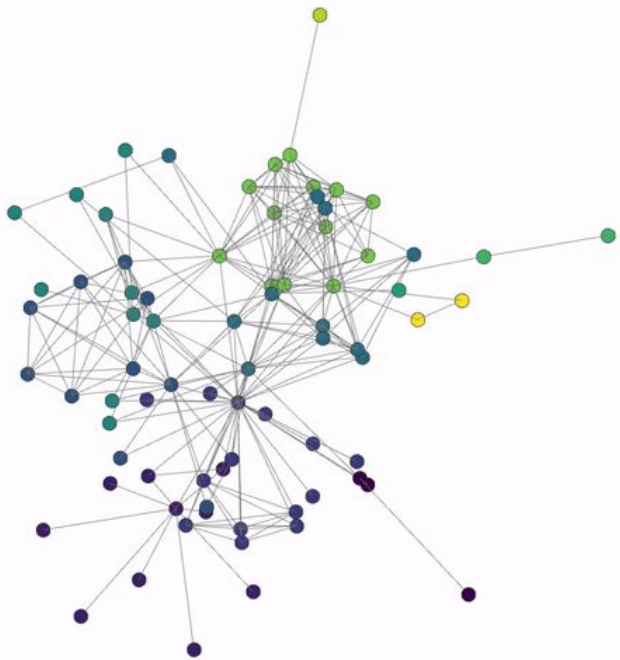
.xml

.yaml

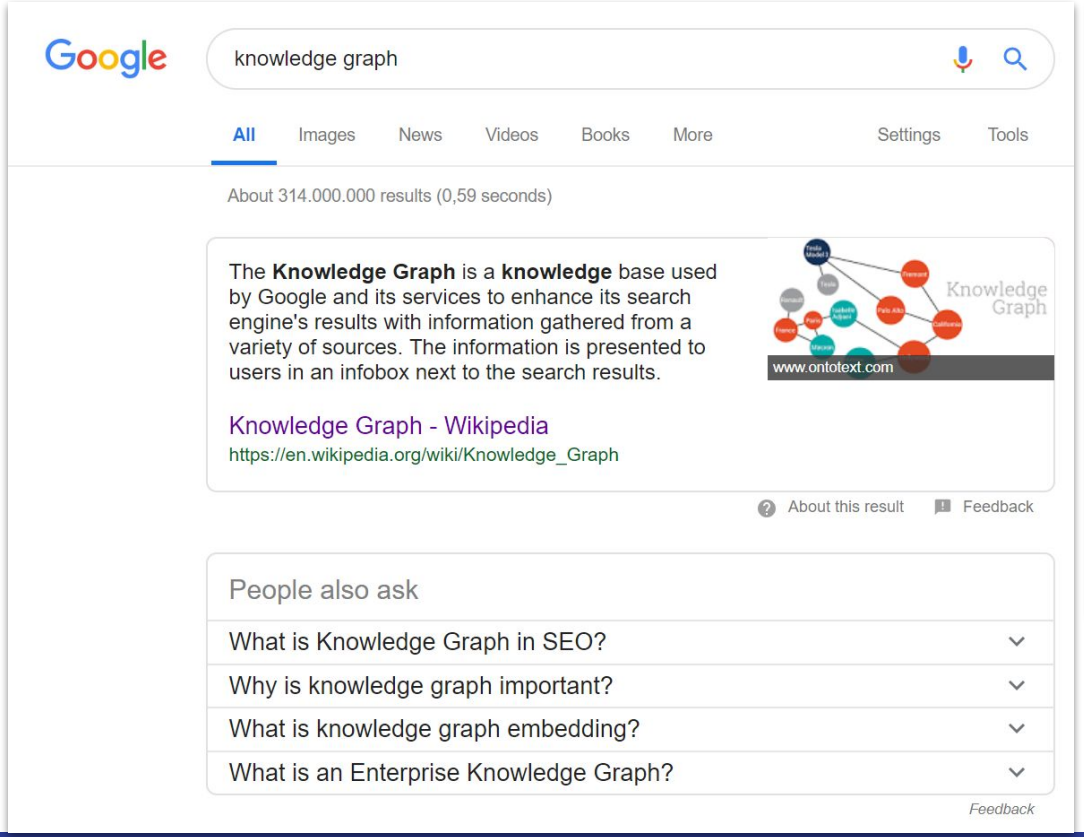
Advantages:
Structured dictionary object. Collection of complex documents with arbitrarily nested data

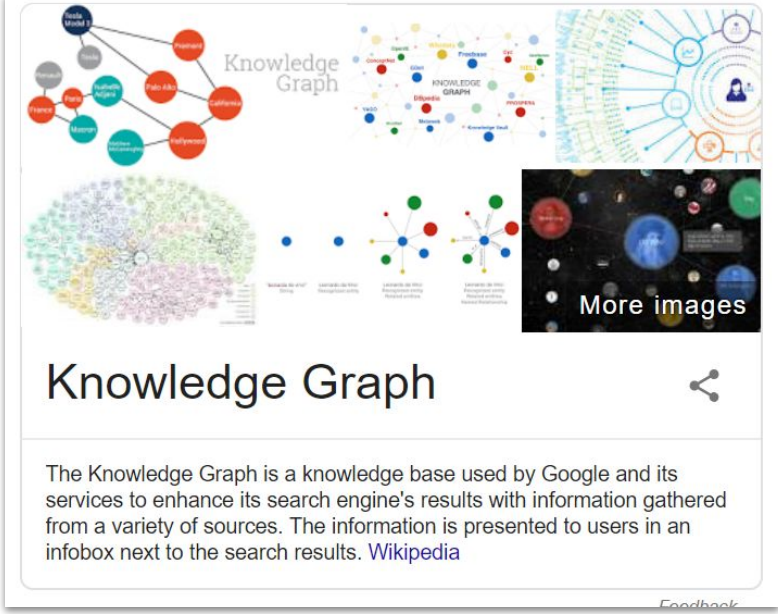
Knowledge Databases


GRAPHS ARE COMING

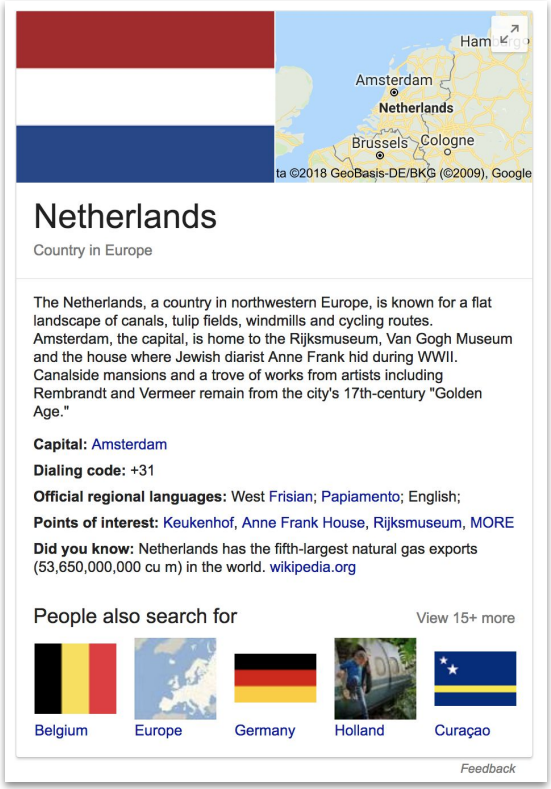


Knowledge Databases = Knowledge Graphs





Knowledge Graphs



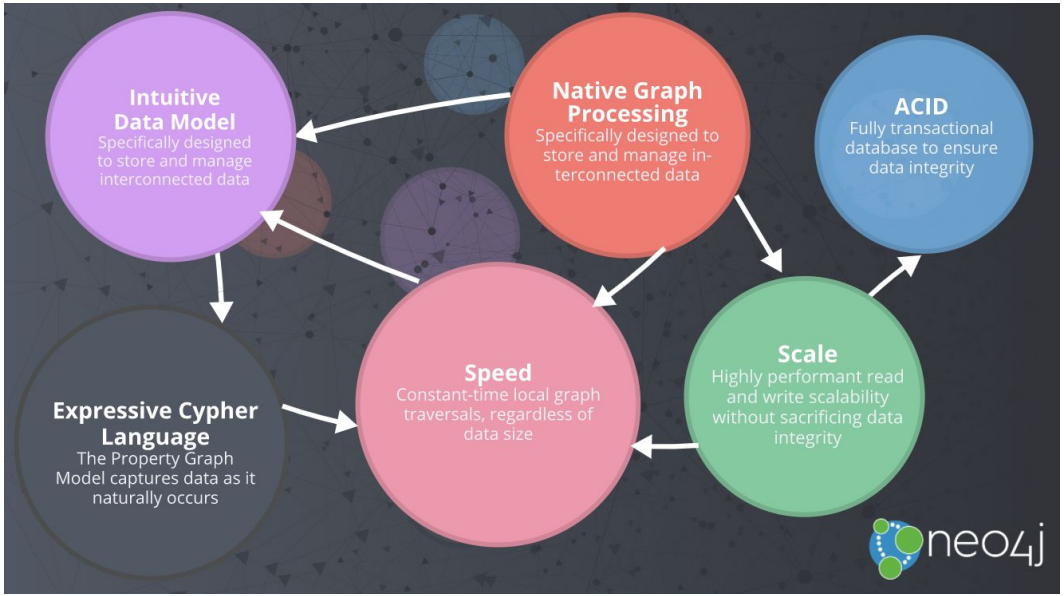
A knowledge graph is a special kind of **database** which **stores knowledge** in a **machine-readable** form and provides a means for information to be collected, organised, shared, searched and utilised.


THE
RENAISSANCE

Knowledge Graphs

- 1956, Rene Bakker (Utrecht) developed the notion of a Knowledge Graph in his PhD as a structure and representation of scientific knowledge.
- Late 1990s endeavours to index web pages using graph analytics techniques.
- Mid-late 2000s, commercial atomicity, consistency, isolation, and durability (ACID) graph databases such as Neo4j and Oracle Spatial and Graph became available.

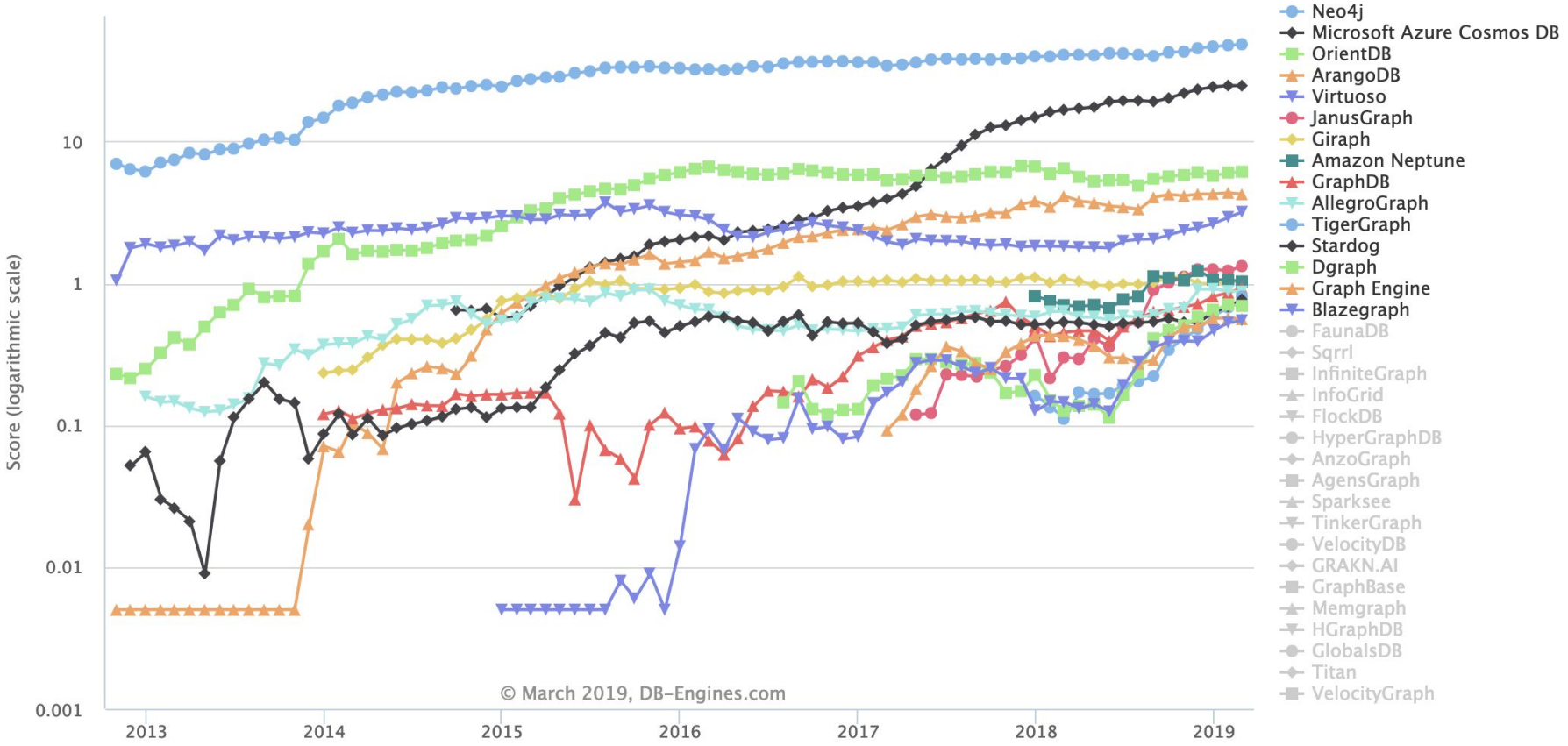
Knowledge Graphs



- Provides a compliant transactional backend for applications.
- Publicly available since 2007. Source code, written in Java and Scala
- The most popular graph-based database
- Cypher query language
- Largest ecosystem

Knowledge Graphs

https://db-engines.com/en/ranking_trend/graph+dbms

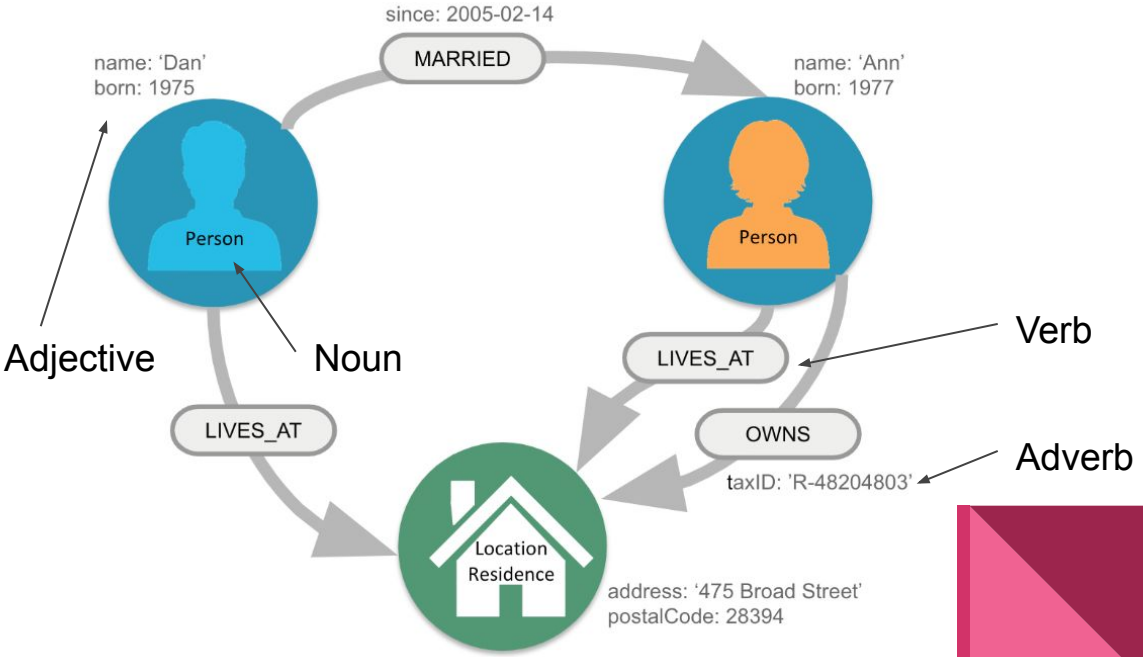


Knowledge Graphs

The graph model should be interpretable as a natural language

- **Noun:** Node label
- **Adjective:** Node property
- **Verb:** Relation label
- **Adverb:** Relation property

Person named ‘Dan’ born in 1975
is MARRIED since 2005-02-14 to
Person named ‘Ann’ born in 1977



Graph Query Language (CYPHER)

( ) **Node**

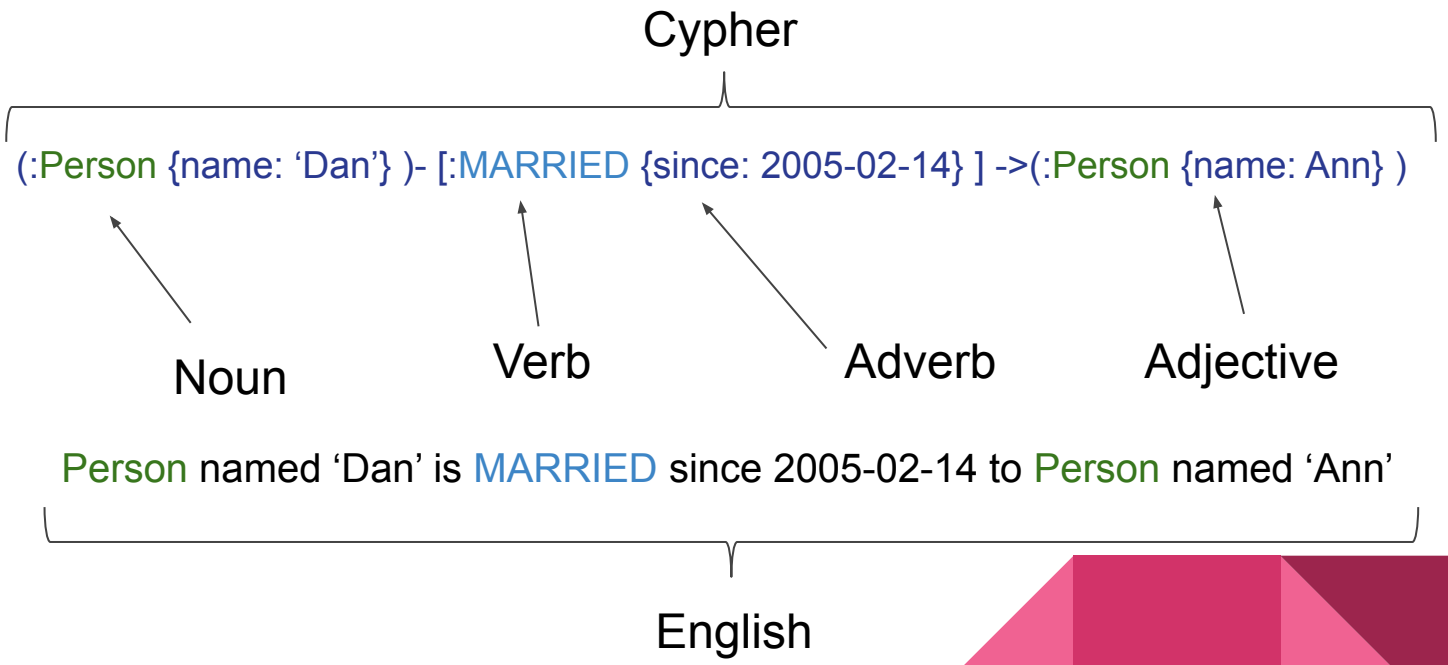
[ ] **Relation**

{ } **Property**

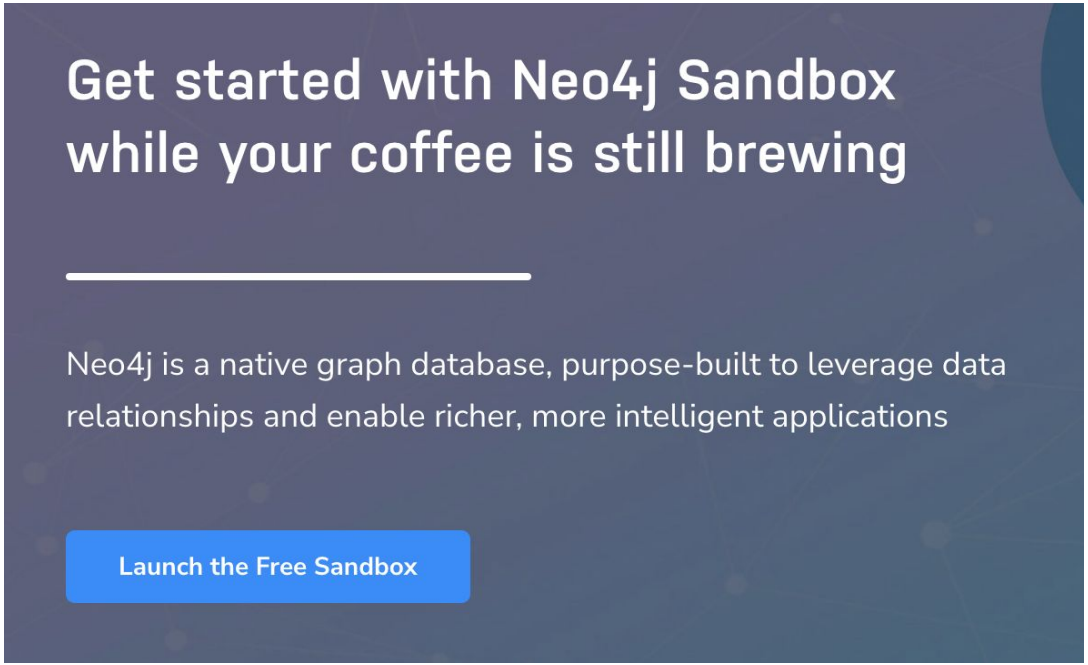
: **Label**

-> **Direction**

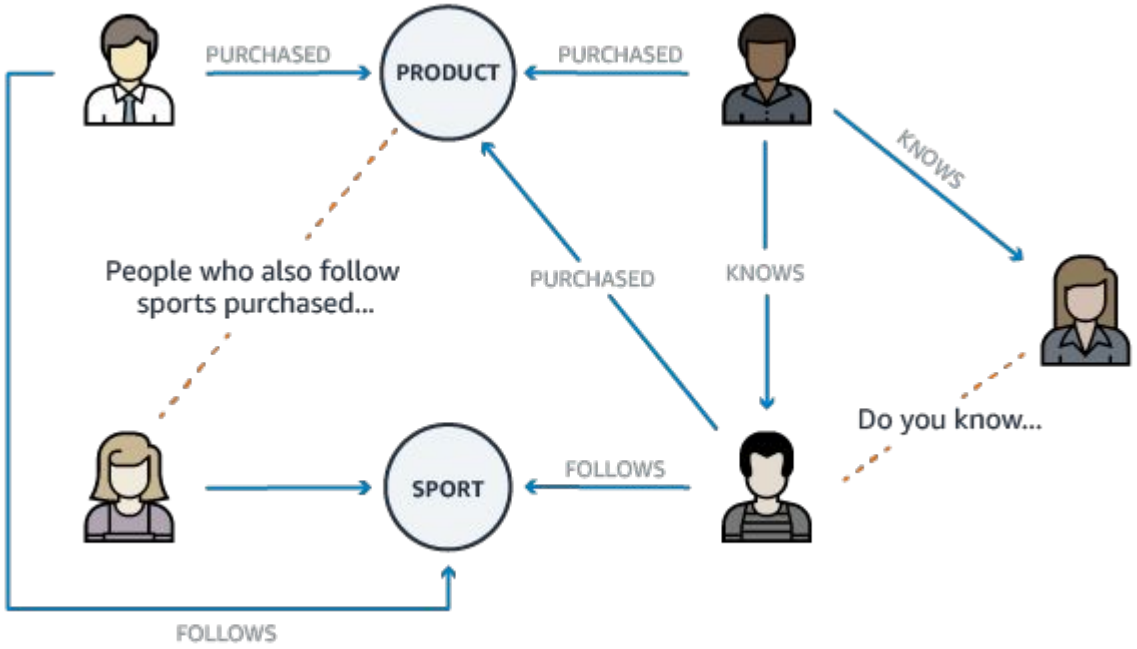
// **Comment**



neo4j.com/sandbox



Applications in Tech



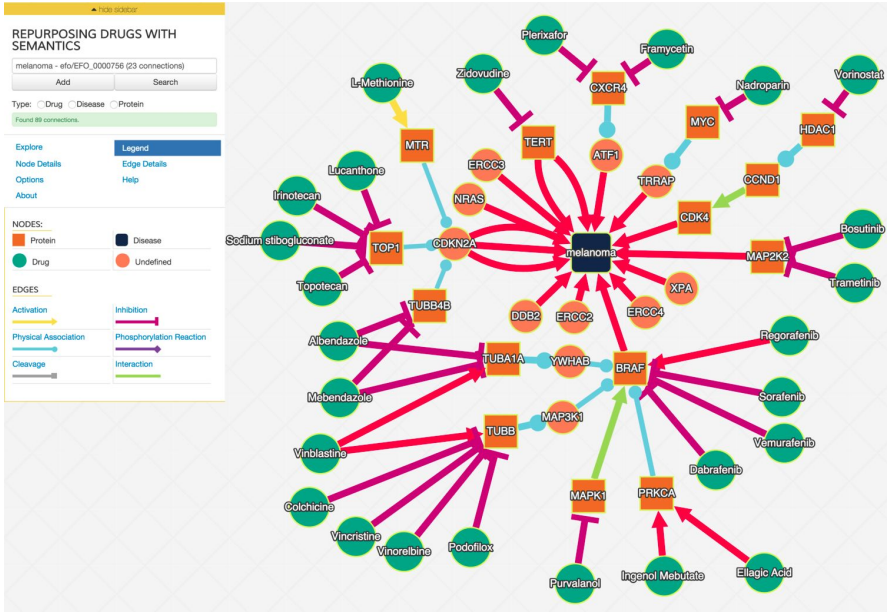
NETFLIX

amazon.com

Applications in Science

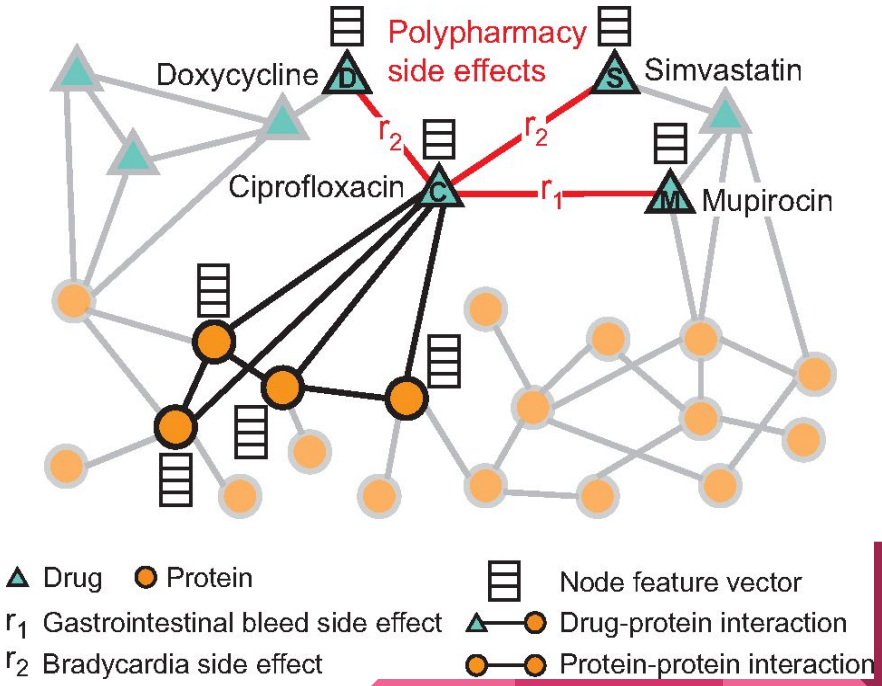
Modeling polypharmacy side effects with graph convolutional networks. Zitnik, Marinka, et al. *Bioinformatics* 34.13 (2018): i457-i466.

Drug discovery



Finding melanoma drugs through a probabilistic knowledge graph. *PeerJ Computer Science. 2017. 3:e106 https://doi.org/10.7717/peerj-cs.106*

Side effects



What are you making?

Search products

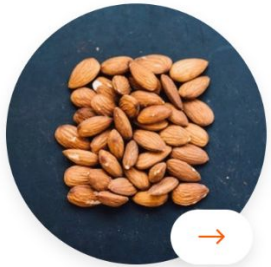
All | Biomass | Ceramic | Chemical | Cosmetic | Feed | Food | Metal | Mineral | Pharmaceutical

Active pharmaceutical

Adhesives

Agar-agar

Alfalfa meal

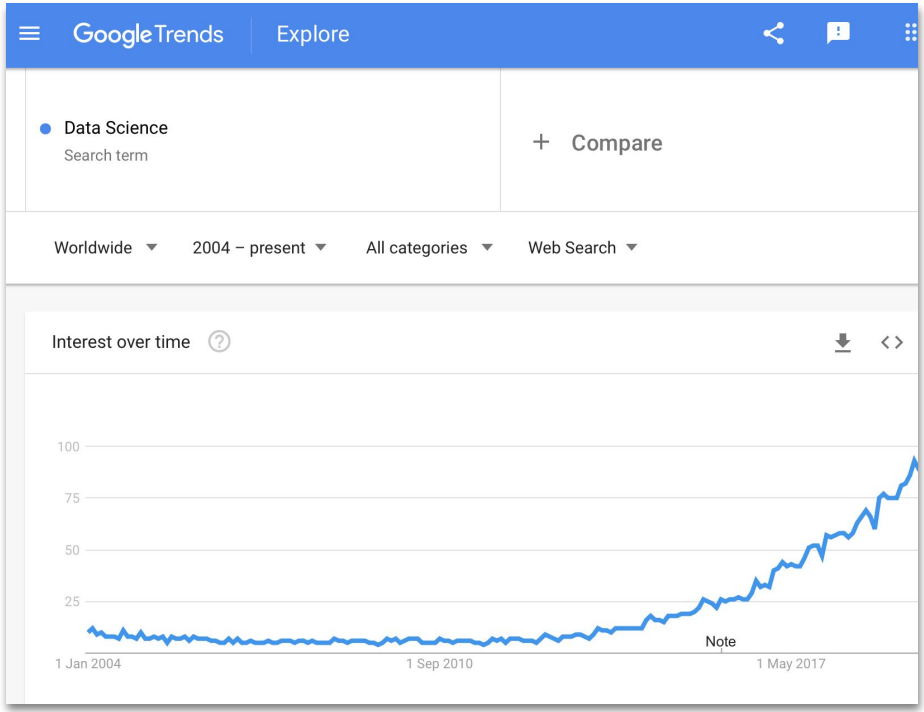


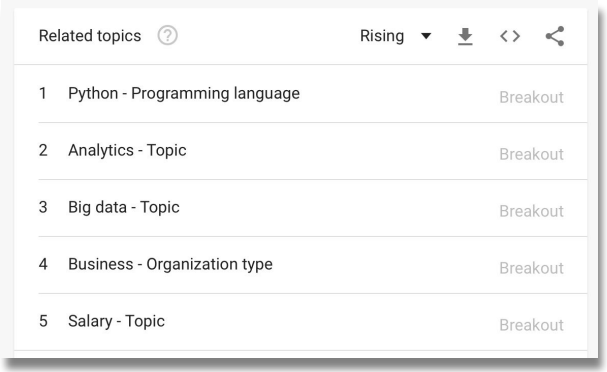
Almonds

What other applications can we think of in knowledge graphs?

Data Science with Python

Is Python for Data Science a Trend?

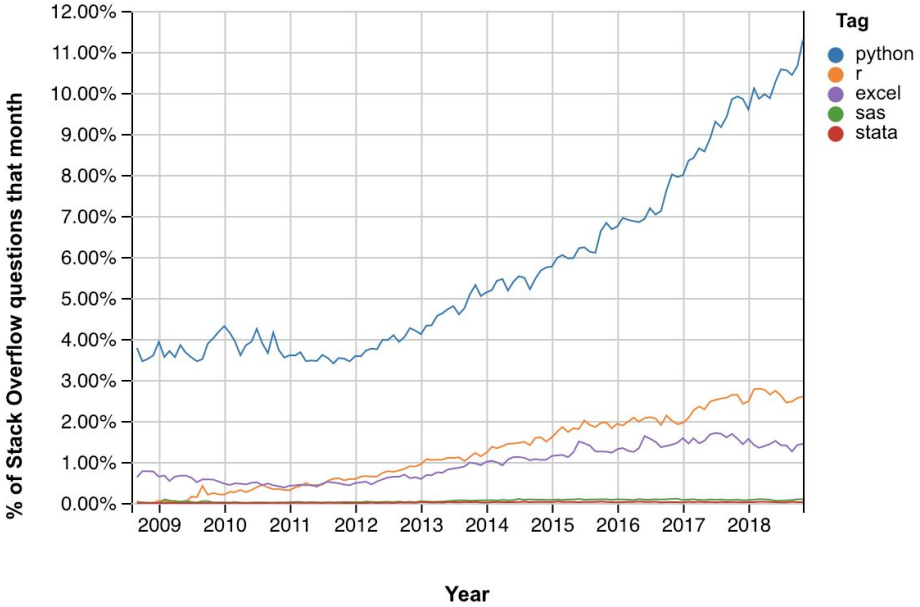


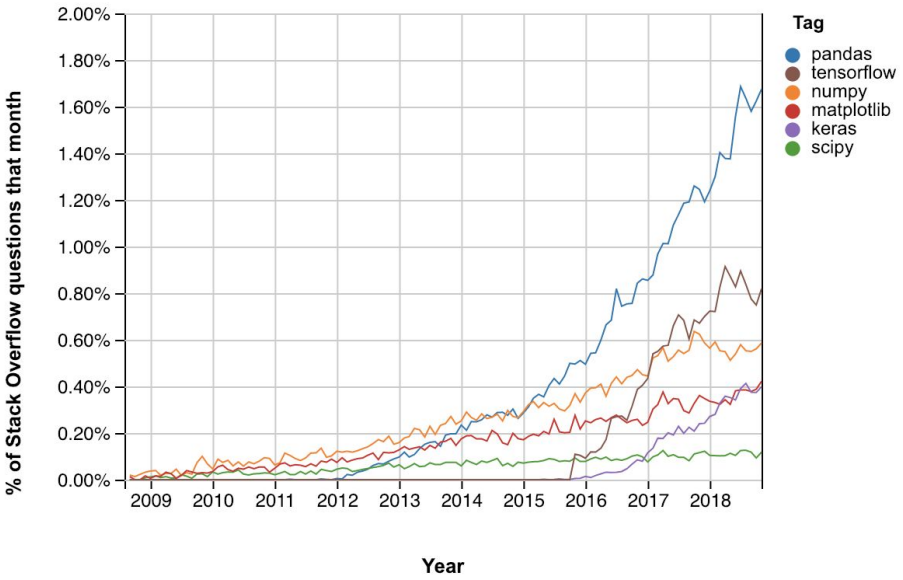


- Data is grows fast. It grows exponentially and continues to grow so. And it's estimated to be about 2.5 Exabytes, that is 2.5 million TB, a day. [1]
- And over 1.5 trillion queries on Google in a year.[2]

[1] https://ourworldindata.org
[2]https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-r
Source: https://insights.stackoverflow.com/trends¶

Data Science with Python





Source: https://insights.stackoverflow.com/trends¶

Data Science with Python

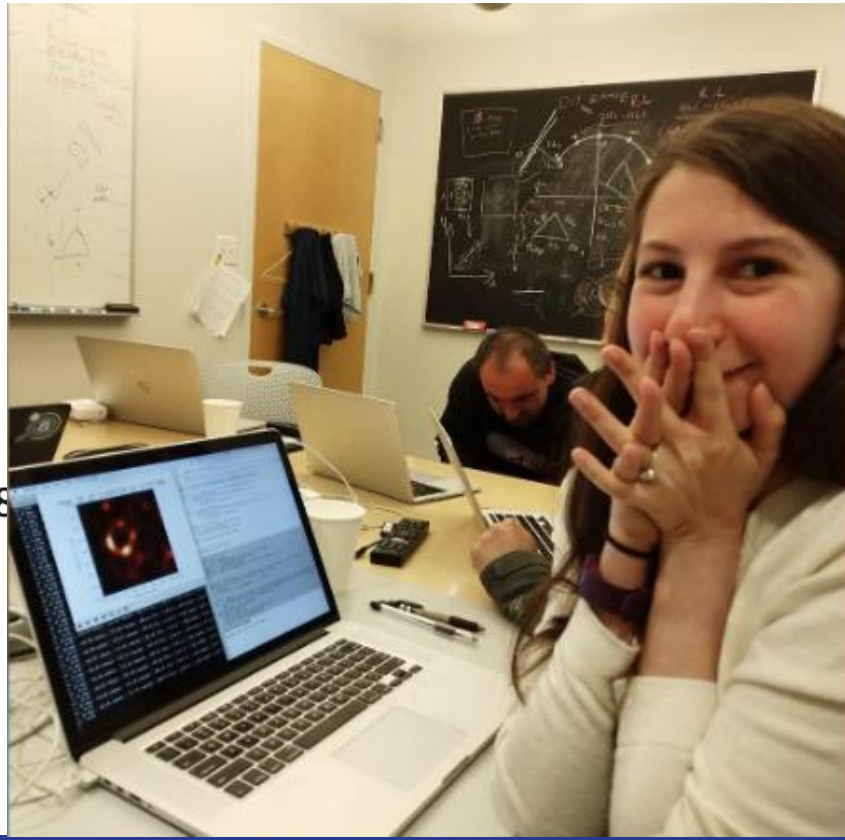
As a general purpose language, Python supports a large range of tasks.
Or put another way: 'Python isn't the best at anything, but it's second best at everything'

- Python is an, open-source high-level, interpreted, programming language.
- A data science/text analytics project may include everything from scraping data from the web, analyzing a mixture or text and numerical data, computing features, training a model, creating high-quality graphs, and then hosting a webapp with results.
- 90,000 libraries in the Python Package Index
- It has a massive user community, who contribute to a large number of high-quality, well maintained open-source tools.
- Widely used in industry.

Source: https://talktotransformer.com

Python brings “sub-technologies”

- Numpy (van der Walt et al. 2011)
- Scipy (Jones et al. 2001)
- Pandas (McKinney 2010)
- Jupyter (Kluyver et al. 2016)
- Matplotlib (Hunter 2007).
- Astropy (The Astropy Collaboration et al. 2013, 2018



Python users

- Google Translate
- Spotify recommender
- Dropbox
- Netflix
- All NLP
- Video games (Sims)
- Visualization software (Dash)

THE ASTROPHYSICAL JOURNAL LETTERS

OPEN ACCESS

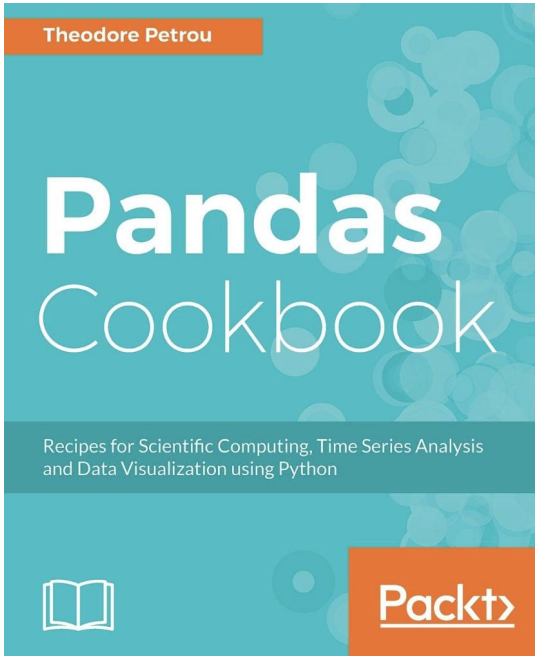
First M87 Event Horizon Telescope Results. III. Data Processing and Calibration

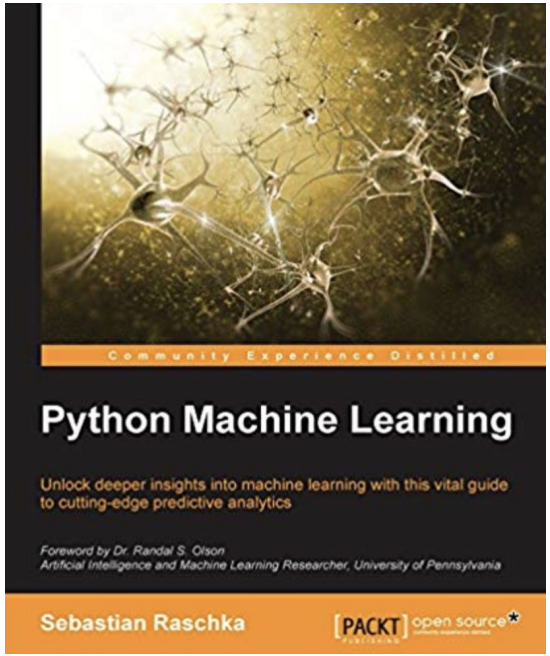
The Event Horizon Telescope Collaboration, Kazunori Akiyama[1,2,3,4], Antxon Alberdi[5], Walter Alef[6], Keiichi Asada[7], Rebecca Azulay[8,9,6], Anne-Kathrin Baczko[6], David Ball[10], Mislav Baloković[4,11], John Barrett[2] +Show full author list

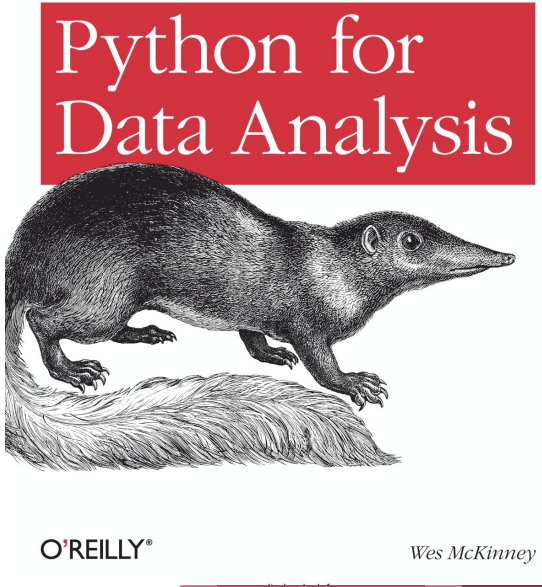
Published 2019 April 10 • © 2019. The American Astronomical Society.
The Astrophysical Journal Letters, Volume 875, Number 1
Focus on the First Event Horizon Telescope Results

It all got started, I believe, because the very earliest Googlers (Sergey, Larry, Craig, ...) made a good engineering decision: "Python where we can, C++ where we must."

Data Science Books







- Notebook 1: bit.ly/stamicarbon1
- Notebook 2: bit.ly/stamicarbon2

- Repo: https://github.com/pedrohserrano/data-science-technologies

- Questionnaire:
- forms.gle/yk5kr57r3UMjZTrL9